PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G10L 11/02

A1

(11) International Publication Number: WO 00/17856

(43) International Publication Date: 30 March 2000 (30.03.00)

(21) International Application Number: PCT/US99/19806

(22) International Filing Date: 27 August 1999 (27.08.99)

(30) Priority Data:
09/156,416 18 September 1998 (18.09.98) US

(71) Applicant: CONEXANT SYSTEMS, INC. [US/US]; Joseph King, 4311 Jamboree Road, Newport Beach, CA 92660–3095 (US).

(72) Inventors: BENYASSINE, Adil; 1305 Reggio Aisle, Irvine, CA 92614 (US). SHLOMOT, Eyal; 86 Costero Aisle, Irvine, CA 92614 (US).

(74) Agent: GESS, Albin, H.; Price, Gess & Ubell, 2100 S.E. Main Street, Suite 250, Irvine, CA 92614 (US).

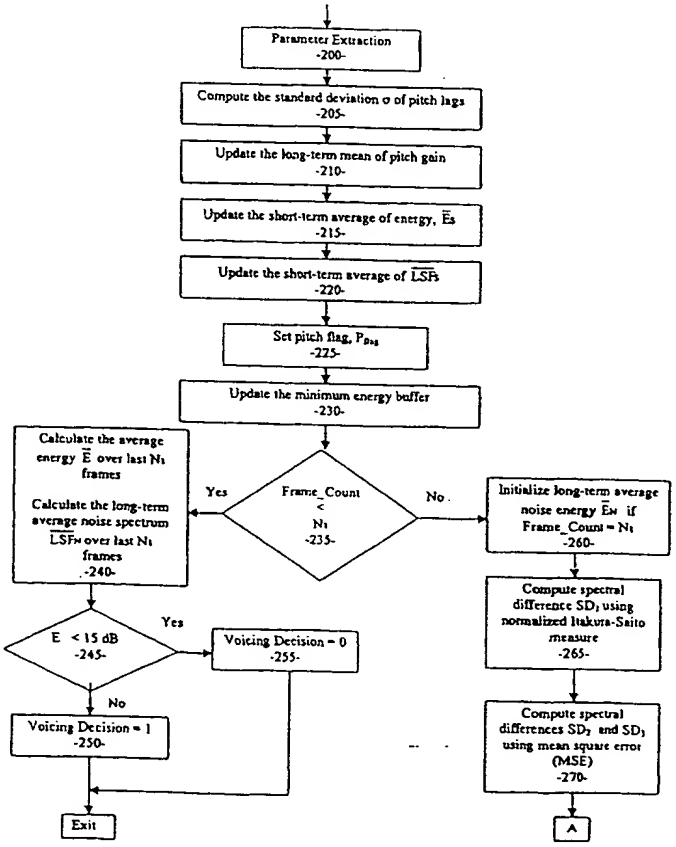
(81) Designated States: CA, CN, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR DETECTING VOICE ACTIVITY IN A SPEECH SIGNAL

(57) Abstract

A method and apparatus for generating frame voicing decisions for an incoming speech signal having periods of active voice and non–active voice for a speech encoder in a speech communications system. A predetermined set of parameters is extracted from the incoming speech signal, including a pitch gain and a pitch lag. A frame voicing decision is made for each frame of the incoming speech signal according to values calculated from the extracted parameters. The predetermined set of parameters further includes a frame full band energy, and a set of spectral parameters called Line Spectral Frequencies (LSF).



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

METHOD AND APPARATUS FOR DETECTING VOICE ACTIVITY IN A SPEECH SIGNAL

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates generally to the field of speech coding in communication systems, and more particularly to detecting voice activity in a communications system.

2. Description of Related Art

Modern communication systems rely heavily on digital speech processing in general, and digital speech compression in particular, in order to provide efficient systems. Examples of such communication systems are digital telephony trunks, voice mail, voice annotation, answering machines, digital voice over data links, etc.

A speech communication system is typically comprised of an encoder, a communication channel and a decoder. At one end of a communications link, the speech encoder converts a speech signal which has been digitized into a bit-stream. The bit-stream is transmitted over the communication channel (which can be a storage medium), and is converted again into a digitized speech signal by the decoder at the other end of the communications link.

The ratio between the number of bits needed for the representation of the digitized speech signal and the number of bits in the bit-stream is the compression ratio. A compression ratio of 12 to 16 is presently achievable, while still maintaining a high quality reconstructed speech signal.

A significant portion of normal speech is comprised of silence, up to an average of 60% during a two-way conversation. During silence, the speech input

device, such as a microphone, picks up the environment or background noise. The noise level and characteristics can vary considerably, from a quiet room to a noisy street or a fast moving car. However, most of the noise sources carry less information than the speech signal and hence a higher compression ratio is achievable during the silence periods. In the following description, speech will be denoted as "active-voice" and silence or background noise will be denoted as "non-active-voice".

The above discussion leads to the concept of dual-mode speech coding schemes, which are usually also variable-rate coding schemes. The active-voice and the non-active voice signals are coded differently in order to improve the system efficiency, thus providing two different modes of speech coding. The different modes of the input signal (active-voice or non-active-voice) are determined by a signal classifier, which can operate external to, or within, the speech encoder. The coding scheme employed for the non-active-voice signal uses less bits and results in an overall higher average compression ratio than the coding scheme employed for the active-voice signal. The classifier output is binary, and is commonly called a "voicing decision." The classifier is also commonly referred to as a Voice Activity Detector ("VAD").

A schematic representation of a speech communication system which employs a VAD for a higher compression rate is depicted in Figure 1. The input to the speech encoder 110 is the digitized incoming speech signal 105. For each frame of a digitized incoming speech signal the VAD 125 provides the voicing decision 140, which is used as a switch 145 between the active-voice encoder 120 and the non-active-voice encoder 115. Either the active-voice bit-stream 135 or the non-active-voice bit-stream 130, together with the voicing decision 140 are transmitted through the communication channel 150. At the speech decoder 155 the voicing decision is used in the switch 160 to select the non-active-voice decoder 165 or the active-voice decoder 170. For each frame, the output of either decoders is used as the reconstructed speech 175.

An example of a method and apparatus which employs such a dual-mode system is disclosed in U.S. Patent No. 5,774,849, commonly assigned to the present

assignee and herein incorporated by reference. According to U.S. Patent No. 5,774,849, four parameters are disclosed which may be used to make the voicing decision. Specifically, the full band energy, the frame low-band energy, a set of parameters called Line Spectral Frequencies ("LSF") and the frame zero crossing rate are compared to a long-term average of the noise signal. While this algorithm provides satisfactory results for many applications, the present inventors have determined that a modified decision algorithm can provide improved performance over the prior art voicing decision algorithms.

SUMMARY OF THE INVENTION

A method and apparatus for generating frame voicing decisions for an incoming speech signal having periods of active voice and non-active voice for a speech encoder in a speech communications system. A predetermined set of parameters is extracted from the incoming speech signal, including a pitch gain and a pitch lag. A frame voicing decision is made for each frame of the incoming speech signal according to values calculated from the extracted parameters. The predetermined set of parameters further includes a frame full band energy, and a set of spectral parameters called Line Spectral Frequencies (LSF).

BRIEF DESCRIPTION OF THE DRAWINGS

The exact nature of this invention, as well as its objects and advantages, will become readily apparent from consideration of the following specification as illustrated in the accompanying drawings, in which like reference numerals designate like parts throughout the figures thereof, and wherein:

Figure 1 is a block diagram representation of a speech communication system using a VAD;

Figures 2(A) and 2(B) are process flowcharts illustrating the operation of the VAD in accordance with the present invention; and

Figure 3 is a block diagram illustrating one embodiment of a VAD according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The following description is provided to enable any person skilled in the art to make and use the invention and sets forth the best modes contemplated by the inventor for carrying out the invention. Various modifications, however, will remain readily apparent to those skilled in the art, since the basic principles of the present invention have been defined herein specifically to provide a voice activity detection method and apparatus.

In the following description, the present invention is described in terms of functional block diagrams and process flow charts, which are the ordinary means for those skilled in the art of speech coding for describing the operation of a VAD. The present invention is not limited to any specific programming languages, or any specific hardware or software implementation, since those skilled in the art can readily determine the most suitable way of implementing the teachings of the present invention.

In the preferred embodiment, a Voice Activity Detection (VAD) module is used to generate a voicing decision which switches between an active-voice encoder/decoder and a non-active-voice encoder/decoder. The binary voicing decision is either 1 (TRUE) for the active-voice or 0 (FALSE) for the non-active-voice.

The VAD process flowchart is illustrated in Figures 2(A) and 2(B). The VAD operates on frames of digitized speech. The frames are processed in time order and are consecutively numbered from the beginning of each conversation/recording. The illustrated process is performed once per frame.

At the first block 200, four parametric features are extracted from the input signal. Extraction of the parameters can be shared with the active-voice encoder module 120 and the non-active-voice encoder module 115 for computational efficiency. The parameters are the frame full band energy, a set of spectral parameters called Line Spectral Frequencies ("LSF"), the pitch gain and the pitch lag. A set of

linear prediction coefficients is derived from the auto correlation and a set of $\{\overline{LSF_i}\}_{i=1}^{p}$ is derived from the set of linear prediction coefficients, as described in ITU-T, Study Group 15 Contribution - Q. 12/15, Draft Recommendation G.729, June 8, 1995, Version 5.0, or DIGITAL SPEECH - Coding for Low Bit Rate Communication Systems by A.M. Kondoz, John Wiley & Son, 1994, England. The full band energy E is the logarithm of the normalized first auto correlation coefficient $R(0)$:

$$E = 10 \cdot \log_{10}\left[\frac{1}{N} R(0)\right],$$

where N is a predetermined normalization factor.

The pitch gain is a measure of the periodicity of the input signal. The higher the pitch gain, the more periodic the signal, and therefore the greater the likelihood that the signal is a speech signal. The pitch lag is the fundamental frequency of the speech (active-voice) signal.

After the parameters are extracted, the standard deviation σ of the pitch lags of the last four previous frames are computed at block 205. The long-term mean of the pitch gain is updated with the average of the pitch gain from the last four frames at block 210. In the preferred embodiment, the long-term mean of the pitch gain is calculated according to the following formula:

$$\overline{P_{gain}} = 0.8 * \overline{P_{gain}} + 0.2 * [\text{average of last four frames}]$$

The short-term average of energy, $\overline{Es}$, is updated at block 215 by averaging the last three frames with the current frame energy. Similarly, the short-term average of LSF vectors, $\overline{LSFs}$, is updated at block 220 by averaging the last three LSF frame vectors with the current LSF frame vector extracted by the parameter extractor at block 200. If the standard deviation σ is less than T_1 or the long-term mean of the pitch gain is greater than T_2, then a flag P_{flag} is set to one, otherwise P_{flag}

equals zero at block 225.

$$\text{If } \sigma < T_1 \text{ OR } P_{gain} > T_2, \text{ then } P_{flag} = 1, \text{ else } P_{flag} = 0.$$

In the preferred embodiment, $T_1 = 1.2$ and $T_2 = 0.7$. At block 230, a minimum energy buffer is updated with the minimum energy value over the last 128 frames. In other words, if the present energy level is less than the minimum energy level determined over the last 128 frames, then the value of the buffer is updated, otherwise the buffer value is unchanged.

If the frame count (i.e. current frame number) is less than a predetermined frame count N_I at block 235, where N_I is 32 in the preferred embodiment, an initialization routine is performed by blocks 240 - 255. At block 240 the average energy $\overline{E}$, and the long-term average noise spectrum $\overline{LSF_N}$ are calculated over the last N_I frames. The average energy $\overline{E}$ is the average of the energy of the last N_I frames. The initial value for $\overline{E}$, calculated at block 240, is:

$$\overline{E} = \frac{1}{N} \sum_{n=1}^{N} E$$

The long-term average noise spectrum $\overline{LSF_N}$ is the average of the LSF vectors of the last N_I frames. At block 245, if the instantaneous energy E extracted at block 200 is less than 15 dB, then the voicing decision is set to zero (block 255), otherwise the voicing decision is set one (block 250). The processing for the frame is then completed and the next frame is processed, beginning with block 200.

The initialization processing of blocks 240-255 initializes the processing over the last few frames. It is not critical to the operation of the present invention and may be skipped. The calculations of block 240 are required, however,

for the proper operation of the invention and should be performed, even if the voicing decisions of blocks 245-255 are skipped. Also, during initialization, the voicing decision could always be set to "1" without significantly impacting the performance of the present invention.

If the frame count is not less than N_1 at block 235, then the first time through block 260 (Frame_Count = N_1), the long-term average noise energy $\overline{E_N}$ is initialized by subtracting 12 dB from the average energy $\overline{E}$:

$$\overline{E_N} = \overline{E} - 12dB$$

Next, at block 265, a spectral difference value SD_1 is calculated using the normalized Itakura-Saito measure. The value SD_1 is a measure of the difference between two spectra (the current frame spectra represented by R and E_π , and the background noise spectrum represented by $\vec{a}$. The Itakurass-Saito measure is a well-known algorithm in the speech processing art and is described in detail, for example, in *Discrete-Time Processing of Speech Signals*, Deller, John R., Proakis, John G. and Hansen, John H.L., 1987, pages 327-329, herein incorporated by reference. Specifically, SD_1 is defined by the following equation:

$$SD_1 = \frac{\vec{a}^T R \vec{a}}{Err}$$

where E_π is the prediction error from linear prediction (LP) analysis of the current frame;

R is the auto-correlation matrix from the LP analysis of the current frame; and

$\vec{a}$ is a linear prediction filter describing the background noise

obtained from $\overline{LSF_N}$.

At block 270 the spectral differences SD_2 and SD_3 are calculated using a mean square error method according to the following equations:

$$SD_2 = \sum_{i=1}^{p} [\overline{LSF_S}(i) - \overline{LSF_N}(i)]^2$$

$$SD_3 = \sum_{i=1}^{p} [\overline{LSF_S}(i) - \overline{LSF}(i)]^2$$

Where $\overline{LSF_S}$ is the short-term average of LSF;

$\overline{LSF_N}$ is the long-term average noise spectrum; and

LSF is the current LSF extracted by the parameter extraction.

The long-term mean of SD_2 (sm_SD_2) in the preferred embodiment is updated at block 275 according to the following equation:

$$sm\_SD_2 = 0.4 * SD_2 + 0.6 * sm\_SD_2$$

Thus, the long term mean of SD_2 is a linear combination of the past long-term mean and the current SD_2 value.

The initial voicing decision, obtained in block 280, is denoted by I_{VD}. The value of I_{VD} is determined according to the following decision statements:

If $\overline{E_S} \geq \overline{E_N} + X1$ dB

OR

$E \rangle \overline{E_N} + X2$ dB

then $IVD = 1$;

If $\overline{E_S} - \overline{E_N} \langle X3$ dB

AND

$sm\_SD_2 \langle T3$

AND

Frame_Count $\rangle$ 128

then $IVD = 0$; else $IVD = 1$;

If $E \rangle 1/2\ (E^{-1} + E^{-2}) + X4$ dB

OR

$SD1 \rangle 1.5$

then $Ivd = 1$.

In the preferred embodiment, $X_1 = 1$, $X_2 = 3$, $X_3 = 2$, $X_4 = 7$, and $T_3 = 0.00012$.

The initial voicing decision is smoothed at block 285 to reflect the long term stationary nature of the speech signal. The smoothed voicing decision of the frame, the previous frame and the frame before the previous frame are denoted by S_{VD}^{0}, S_{VD}^{-1} and S_{VD}^{-2}, respectively. Both S_{VD}^{-1} and S_{VD}^{-2} are initialized to 1 and $S_{VD}^{0} = I_{VD}$. A Boolean parameter F_{VD}^{-1} is initialized to 1 and a counter denoted by C_e is initialized to 0. The energy of the previous frame is denoted by E_{-1}. Thus, the smoothing stage is defined by:

$$
\begin{aligned}
&\text{if } F_{VD}^{-1} = 1 \text{ and } I_{VD} = 0 \text{ and } S_{VD}^{-1} = 1 \text{ and } S_{VD}^{-2} = 1 \\
&\quad S_{VD}^{0} = 1 \\
&\quad C_e = C_e + 1 \\
&\quad \text{if } C_e \le T_4 \ \{ \\
&\qquad F_{VD}^{-1} = 1 \\
&\quad \} \\
&\quad \text{else } \{ \\
&\qquad F_{VD}^{-1} = 0 \\
&\qquad C_e = 0 \\
&\quad \} \\
&\} \\
&\text{else} \\
&\quad F_{VD}^{-1} = 1
\end{aligned}
$$

Ce is reset to 0 if $S_{VD}^{-1} = 1$ and $S_{VD}^{-2} = 1$ and $I_{VD} = 1$.

If $P_{flag} = 1$, then $S^{0}_{VD} = 1$

If E < 15 dB, then $S^{0}_{VD} = 0$

In the preferred embodiment, $T_4 = 14$. The final value of S^{0}_{VD} represents the final voicing decision, with a value of "1" representing an active voice speech signal, and a value of "0" representing a non-active voice speech signal.

F_{SD} is a flag which indicates whether consecutive frames exhibit spectral stationarity (i.e., spectrum does not change dramatically from frame to frame). F_{SD} is set at block 290 according to the following where C_S is a counter initialized to 0.

If Frame_Count > 128 AND $SD_3 < T_5$
then
$C_S = C_S + 1$
else
$C_S = 0$;
If $C_S > N$
$F_{SD} = 1$
else
$F_{SD} = 0$.

In the preferred embodiment, T5 = 0.0005 and N = 20.

The running averages of the background noise characteristics are updated at the last stage of the VAD algorithm. At block 295 and 300, the following conditions are tested and the updating takes place only if these conditions are met:

If $\bar{E}_S < \bar{E}_N + 3$ AND $P_{flag} = 0$
then $E_N = \beta_{EN} * \overline{E_N} + (1 - \beta_{EN}) * [\max \text{ of } E \text{ AND } \overline{E_S}]$
AND
$\overline{LSF_N}(i) = \beta_{LSF} * \overline{LSF_N}(i) + (1 - \beta_{LSF}) * LSF(i)$ $i = 1, \ldots p$
If Frame_Count > 128 AND
$\bar{E}_N < \text{Min}$ AND $F_{SD} = 1$ AND $P_{flag} = 0$
then
$\bar{E}_N = \text{Min}$
else If Frame_Count > 128 AND $\bar{E}_N > \text{Min} + 10$
then
$\overline{E_N} = \text{Min}$.

Figure 3 illustrates a block diagram of one possible implementation of a VAD 400 according to the present invention. An extractor 402 extracts the required predetermined parameters, including a pitch lag and a pitch gain, from the incoming

speech signal 105. A calculator unit 404 performs the necessary calculations on the extracted parameters, as illustrated by the flowcharts in Figs. 2(A) and 2(B). A decision unit 406 then determines whether a current speech frame is an active voice or a non-active voice signal and outputs a voicing decision 140 (as shown in Fig. 1).

Those skilled in the art will appreciate that various adaptations and modifications of the just-described preferred embodiments can be configured without departing from the scope and spirit of the invention. Therefore, it is to be understood that within the scope of the appended claims, the invention may be practiced other than as specifically described herein.

CLAIMS

What Is Claimed Is:

1. In a speech communication system, a method for generating a frame voicing decision comprising the steps of:

(a) extracting a predetermined set of parameters, including a pitch gain and a pitch lag, from the incoming speech signal for each frame; and

(b) making a frame voicing decision according to the extracted predetermined set of parameters.

2. The method according to claim 1, wherein the predetermined set of parameters further comprises a full band energy and line spectral frequencies (LSF).

3. A method according to claim 2, wherein the step of making a frame voicing decision further comprises the steps of:

i. calculating a standard deviation σ of the pitch lag;
ii. calculating a long-term mean of pitch gain;
iii. calculating a short-term average of energy E, $\overline{E}_S$;
iv. calculating a short-term average of $\overline{LSF}_S$;
v. calculating an average energy $\overline{E}$; and
vi. calculating an average LSF value, $\overline{LSF}_N$.

4. A method according to claim 3, wherein the step of making a frame voicing decision further comprises the steps of:

i) calculating a spectral difference SD_1 using a normalized Itakura-Saito measure;

ii) calculating a spectral difference SD_2 using a mean square error method;

iii) calculating a spectral difference SD_3 using a mean square error method; and

iv) calculating a long-term mean of SD_2.

5. A method according to claim 4, wherein an initial frame voicing decision is made according to the calculated values.

6. A method according to claim 5, wherein the initial frame voicing decision is smoothed.

7. A method according to claim 6, wherein an initialization routine is performed for a predetermined number of initial frames, such that the voicing decision is set to active voice.

8. A voice activity detector (VAD) for making a voicing decision on an incoming speech signal frame, the VAD comprising:

an extractor for extracting a predetermined set of parameters, including a pitch gain and a pitch lag, from the incoming speech signal for each frame;

a calculator unit for calculating a set of predetermined values based on the extracted predetermined set of parameters: and

a decision unit for making a frame voicing decision according to the predetermined set of values.

9. The VAD according to claim 8, wherein the predetermined set of parameters further comprises a full band energy and line spectral frequencies (LSF).

10. The VAD according to claim 9, wherein the calculator unit calculates:

a standard deviation σ of the pitch lag;

a long-term mean of pitch gain;

a short-term average of energy E, $\bar{E}s$;

a short-term average of LSF, $\overline{LSFs}$;

an average energy $\bar{E}$; and

an average LSF value. $\overline{LSF_N}$.

11. The VAD according to claim 10, wherein the calculator unit further calculates:

a spectral difference SD, using a normalized Itakura-Saito

measure:

a spectral difference SD_2 using a mean square error method;

a spectral difference SD_3 using a mean square error method; and

a long-term mean of SD_2.

12. The VAD according to claim 11, wherein the decision unit makes an initial frame voicing decision according to the values calculated by the calculation means

13. The VAD according to claim 12, wherein the initial frame voicing decision is smoothed.

14. A voice activity detection method for detecting voice activity in an incoming speech signal frame, the improvement comprising making a voicing decision based on a pitch lag and a pitch gain of the speech signal frame.

15. The voice activity detection method of claim 14, further comprising making the voicing decision based on a frame full band energy and a set of spectral parameters called Line Spectral Frequencies (LSF).

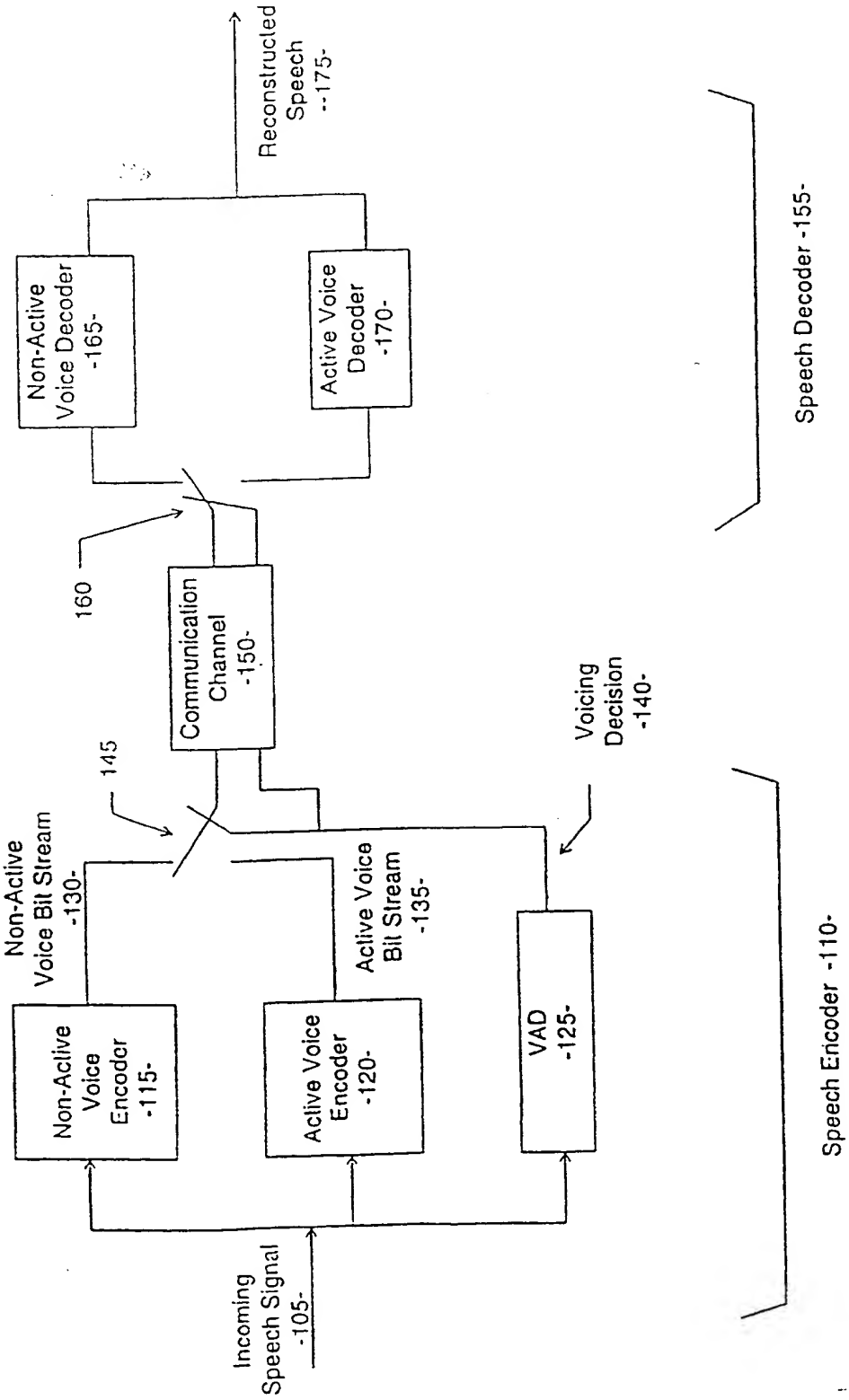


Figure 1 Prior Art

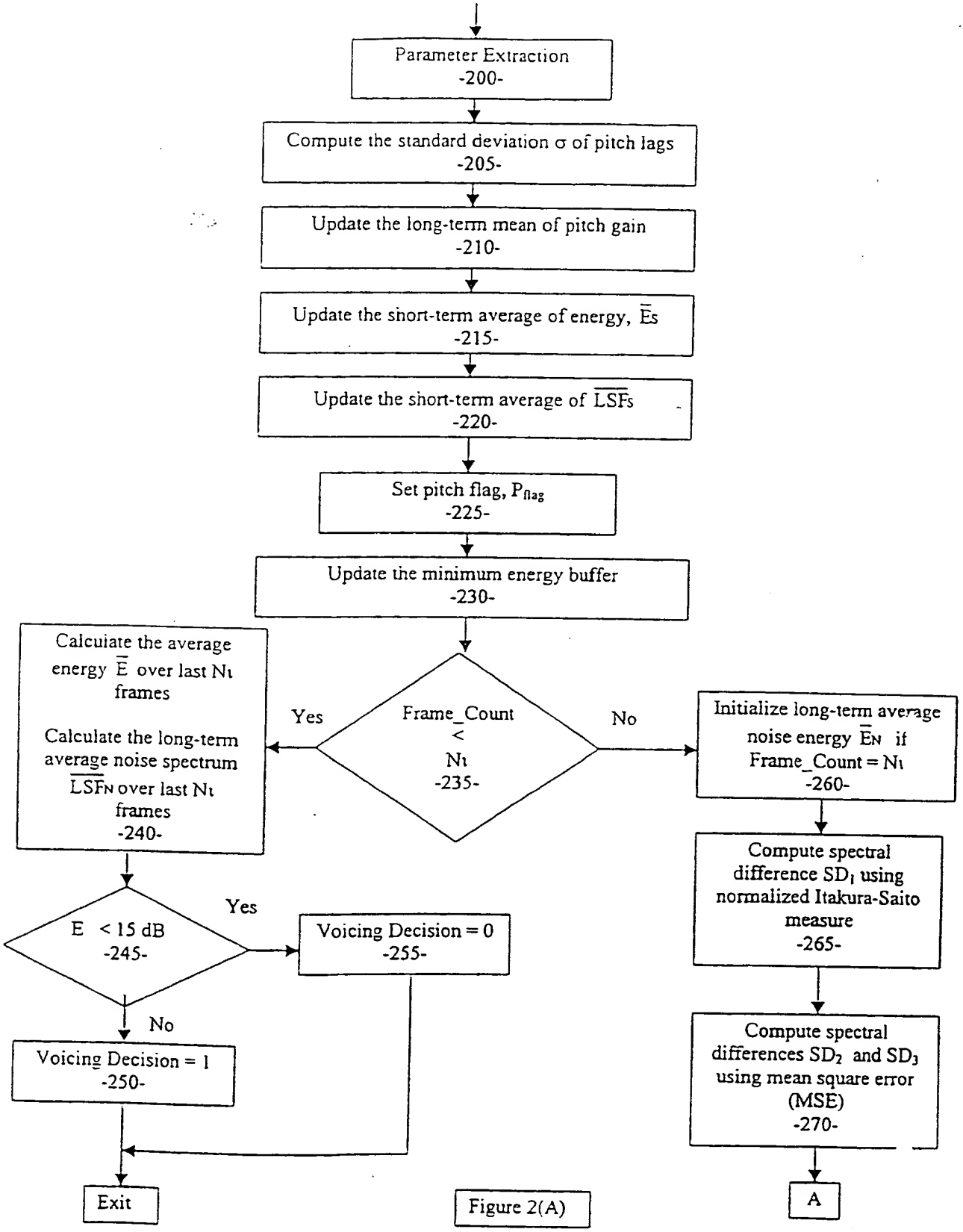

Parameter Extraction
-200-
Compute the standard deviation σ of pitch lags
-205-
Update the long-term mean of pitch gain
-210-
Update the short-term average of energy, $\bar{E}_S$
-215-
Update the short-term average of $\overline{LSF}_S$
-220-
Set pitch flag, P_{flag}
-225-
Update the minimum energy buffer
-230-
Frame_Count < N_1
-235-
Yes
No
Calculate the average energy $\bar{E}$ over last N_1 frames
Calculate the long-term average noise spectrum $\overline{LSF}_N$ over last N_1 frames
-240-
Initialize long-term average noise energy $\bar{E}_N$ if Frame_Count = N_1
-260-
E < 15 dB
-245-
Yes
No
Voicing Decision = 0
-255-
Voicing Decision = 1
-250-
Compute spectral difference SD_1 using normalized Itakura-Saito measure
-265-
Compute spectral differences SD_2 and SD_3 using mean square error (MSE)
-270-
Exit
A

Figure 2(A)

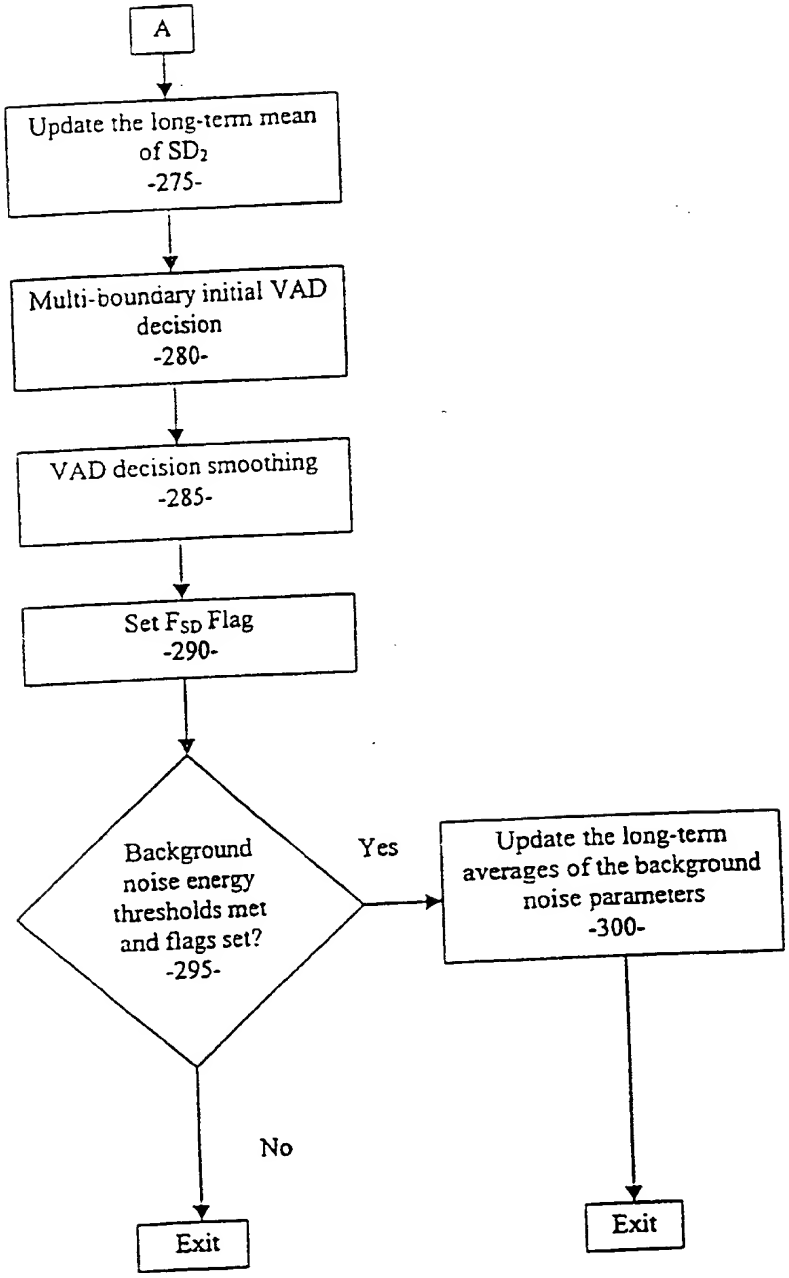

A
Update the long-term mean of SD_2
-275-
Multi-boundary initial VAD decision
-280-
VAD decision smoothing
-285-
Set F_{SD} Flag
-290-
Background noise energy thresholds met and flags set?
-295-
Yes
Update the long-term averages of the background noise parameters
-300-
No
Exit
Exit

Figure 2(B)

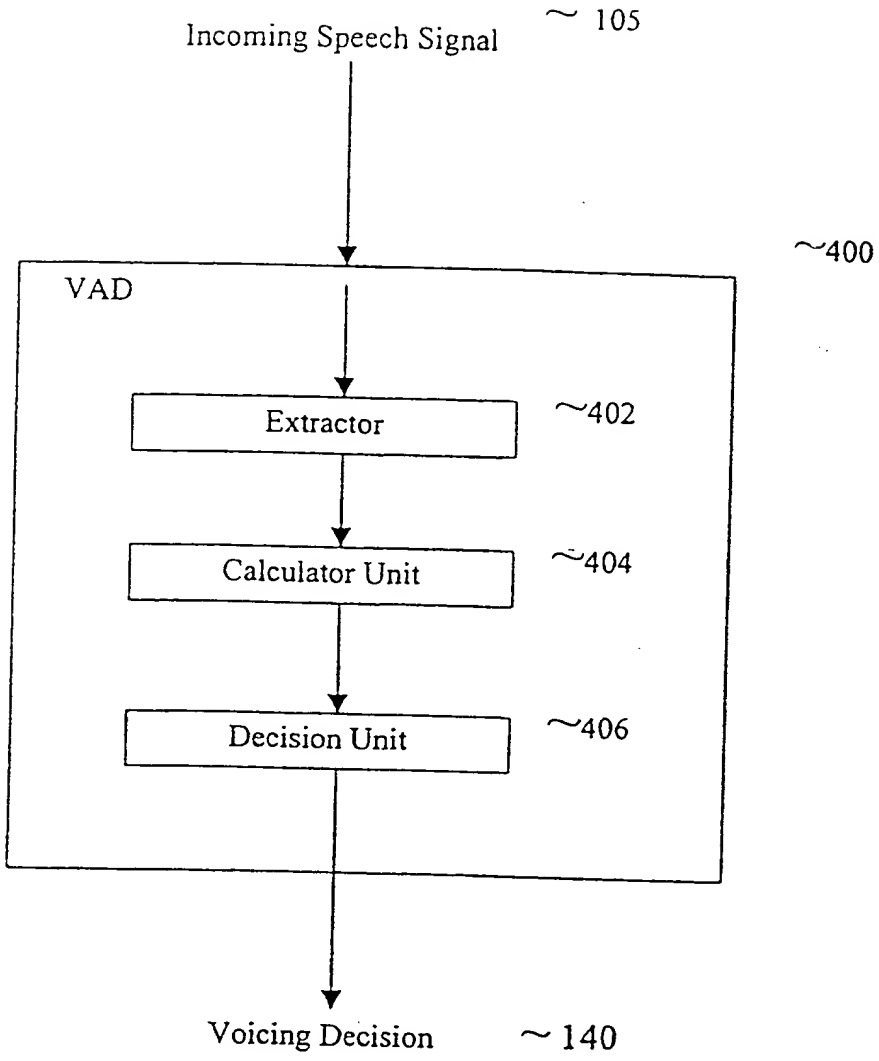

Incoming Speech Signal
105
400
VAD
Extractor
402
Calculator Unit
404
Decision Unit
406
Voicing Decision
140

Figure 3

INTERNATIONAL SEARCH REPORT

In Sonal Application No
PCT/US 99/19806

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L11/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 784 311 A (NOKIA MOBILE PHONES LTD) 16 July 1997 (1997-07-16) abstract; figure 2 page 8, line 15 - line 20	1,8,14
Y		2-4, 9-11,15
Y	EP 0 785 419 A (ROCKWELL INTERNATIONAL CORP) 23 July 1997 (1997-07-23) cited in the application claims 1-14	2-4, 9-11,15

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 December 1999	11/01/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax (+31–70) 340–3016	Van Doremalen, J

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Ir rtional Application No
PCT/US 99/19806

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0784311 A	16-07-1997	FI 955947 A	13-06-1997
		AU 1067797 A	03-07-1997
		AU 1067897 A	03-07-1997
		EP 0790599 A	20-08-1997
		WO 9722116 A	19-06-1997
		WO 9722117 A	19-06-1997
		JP 9212195 A	15-08-1997
		JP 9204196 A	05-08-1997
		US 5839101 A	17-11-1998
		US 5963901 A	05-10-1999
EP 0785419 A	23-07-1997	US 5774849 A	30-06-1998
		JP 9198099 A	31-07-1997

VERSION*

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G10L 11/02

A1

(11) International Publication Number: WO 00/17856

(43) International Publication Date: 30 March 2000 (30.03.00)

(21) International Application Number: PCT/US99/19806

(22) International Filing Date: 27 August 1999 (27.08.99)

(30) Priority Data:
09/156.416 18 September 1998 (18.09.98) US

(71) Applicant: CONEXANT SYSTEMS, INC. [US/US]; Joseph King, 4311 Jamboree Road, Newport Beach, CA 92660-3095 (US).

(72) Inventors: BENYASSINE, Adil; 1305 Reggio Aisle, Irvine, CA 92614 (US). SHLOMOT, Eyal; 86 Costero Aisle, Irvine, CA 92614 (US).

(74) Agent: GESS, Albin, H.; Price, Gess & Ubell, 2100 S.E. Main Street, Suite 250, Irvine, CA 92614 (US).

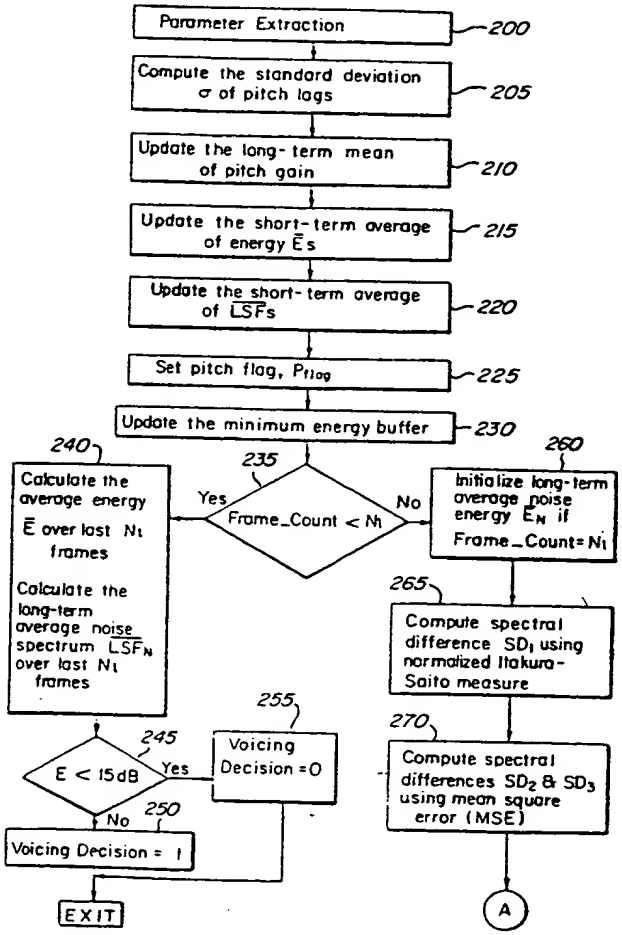
(81) Designated States: CA, CN, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR DETECTING VOICE ACTIVITY IN A SPEECH SIGNAL

(57) Abstract

A method and apparatus for generating frame voicing decisions for an incoming speech signal having periods of active voice and non-active voice for a speech encoder in a speech communications system. A predetermined set of parameters is extracted from the incoming speech signal, including a pitch gain and a pitch lag. A frame voicing decision is made for each frame of the incoming speech signal according to values calculated from the extracted parameters. The predetermined set of parameters further includes a frame full band energy, and a set of spectral parameters called Line Spectral Frequencies (LSF).



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

METHOD AND APPARATUS FOR DETECTING VOICE ACTIVITY IN A SPEECH SIGNAL

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates generally to the field of speech coding in communication systems, and more particularly to detecting voice activity in a communications system.

2. Description of Related Art

Modern communication systems rely heavily on digital speech processing in general, and digital speech compression in particular, in order to provide efficient systems. Examples of such communication systems are digital telephony trunks, voice mail, voice annotation, answering machines, digital voice over data links, etc.

A speech communication system is typically comprised of an encoder, a communication channel and a decoder. At one end of a communications link, the speech encoder converts a speech signal which has been digitized into a bit-stream. The bit-stream is transmitted over the communication channel (which can be a storage medium), and is converted again into a digitized speech signal by the decoder at the other end of the communications link.

The ratio between the number of bits needed for the representation of the digitized speech signal and the number of bits in the bit-stream is the compression ratio. A compression ratio of 12 to 16 is presently achievable, while still maintaining a high quality reconstructed speech signal.

A significant portion of normal speech is comprised of silence, up to an average of 60% during a two-way conversation. During silence, the speech input

device, such as a microphone, picks up the environment or background noise. The noise level and characteristics can vary considerably, from a quiet room to a noisy street or a fast moving car. However, most of the noise sources carry less information than the speech signal and hence a higher compression ratio is achievable during the silence periods. In the following description, speech will be denoted as "active-voice" and silence or background noise will be denoted as "non-active-voice".

The above discussion leads to the concept of dual-mode speech coding schemes, which are usually also variable-rate coding schemes. The active-voice and the non-active voice signals are coded differently in order to improve the system efficiency, thus providing two different modes of speech coding. The different modes of the input signal (active-voice or non-active-voice) are determined by a signal classifier, which can operate external to, or within, the speech encoder. The coding scheme employed for the non-active-voice signal uses less bits and results in an overall higher average compression ratio than the coding scheme employed for the active-voice signal. The classifier output is binary, and is commonly called a "voicing decision." The classifier is also commonly referred to as a Voice Activity Detector ("VAD").

A schematic representation of a speech communication system which employs a VAD for a higher compression rate is depicted in Figure 1. The input to the speech encoder 110 is the digitized incoming speech signal 105. For each frame of a digitized incoming speech signal the VAD 125 provides the voicing decision 140, which is used as a switch 145 between the active-voice encoder 120 and the non-active-voice encoder 115. Either the active-voice bit-stream 135 or the non-active-voice bit-stream 130, together with the voicing decision 140 are transmitted through the communication channel 150. At the speech decoder 155 the voicing decision is used in the switch 160 to select the non-active-voice decoder 165 or the active-voice decoder 170. For each frame, the output of either decoders is used as the reconstructed speech 175.

An example of a method and apparatus which employs such a dual-mode system is disclosed in U.S. Patent No. 5,774,849, commonly assigned to the present

assignee and herein incorporated by reference. According to U.S. Patent No. 5,774,849, four parameters are disclosed which may be used to make the voicing decision. Specifically, the full band energy, the frame low-band energy, a set of parameters called Line Spectral Frequencies ("LSF") and the frame zero crossing rate are compared to a long-term average of the noise signal. While this algorithm provides satisfactory results for many applications, the present inventors have determined that a modified decision algorithm can provide improved performance over the prior art voicing decision algorithms.

SUMMARY OF THE INVENTION

A method and apparatus for generating frame voicing decisions for an incoming speech signal having periods of active voice and non-active voice for a speech encoder in a speech communications system. A predetermined set of parameters is extracted from the incoming speech signal, including a pitch gain and a pitch lag. A frame voicing decision is made for each frame of the incoming speech signal according to values calculated from the extracted parameters. The predetermined set of parameters further includes a frame full band energy, and a set of spectral parameters called Line Spectral Frequencies (LSF).

BRIEF DESCRIPTION OF THE DRAWINGS

The exact nature of this invention, as well as its objects and advantages, will become readily apparent from consideration of the following specification as illustrated in the accompanying drawings, in which like reference numerals designate like parts throughout the figures thereof, and wherein:

Figure 1 is a block diagram representation of a speech communication system using a VAD;

Figures 2(A) and 2(B) are process flowcharts illustrating the operation of the VAD in accordance with the present invention; and

Figure 3 is a block diagram illustrating one embodiment of a VAD according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The following description is provided to enable any person skilled in the art to make and use the invention and sets forth the best modes contemplated by the inventor for carrying out the invention. Various modifications, however, will remain readily apparent to those skilled in the art, since the basic principles of the present invention have been defined herein specifically to provide a voice activity detection method and apparatus.

In the following description, the present invention is described in terms of functional block diagrams and process flow charts, which are the ordinary means for those skilled in the art of speech coding for describing the operation of a VAD. The present invention is not limited to any specific programming languages, or any specific hardware or software implementation, since those skilled in the art can readily determine the most suitable way of implementing the teachings of the present invention.

In the preferred embodiment, a Voice Activity Detection (VAD) module is used to generate a voicing decision which switches between an active-voice encoder/decoder and a non-active-voice encoder/decoder. The binary voicing decision is either 1 (TRUE) for the active-voice or 0 (FALSE) for the non-active-voice.

The VAD process flowchart is illustrated in Figures 2(A) and 2(B). The VAD operates on frames of digitized speech. The frames are processed in time order and are consecutively numbered from the beginning of each conversation/recording. The illustrated process is performed once per frame.

At the first block 200, four parametric features are extracted from the input signal. Extraction of the parameters can be shared with the active-voice encoder module 120 and the non-active-voice encoder module 115 for computational efficiency. The parameters are the frame full band energy, a set of spectral parameters called Line Spectral Frequencies ("LSF"), the pitch gain and the pitch lag. A set of

linear prediction coefficients is derived from the auto correlation and a set of $\{\overline{LSF_i}\}_{i=1}^{p}$ is derived from the set of linear prediction coefficients, as described in ITU-T, Study Group 15 Contribution - Q. 12/15, Draft Recommendation G.729, June 8, 1995, Version 5.0, or DIGITAL SPEECH - Coding for Low Bit Rate Communication Systems by A.M. Kondoz, John Wiley & Son, 1994, England. The full band energy E is the logarithm of the normalized first auto correlation coefficient $R(0)$:

$$E = 10 \cdot \log_{10}\left[\frac{1}{N} R(0)\right],$$

where N is a predetermined normalization factor.

The pitch gain is a measure of the periodicity of the input signal. The higher the pitch gain, the more periodic the signal, and therefore the greater the likelihood that the signal is a speech signal. The pitch lag is the fundamental frequency of the speech (active-voice) signal.

After the parameters are extracted, the standard deviation σ of the pitch lags of the last four previous frames are computed at block 205. The long-term mean of the pitch gain is updated with the average of the pitch gain from the last four frames at block 210. In the preferred embodiment, the long-term mean of the pitch gain is calculated according to the following formula:

$$\overline{P_{gain}} = 0.8 * \overline{P_{gain}} + 0.2 * [\text{average of last four frames}]$$

The short-term average of energy, $\overline{Es}$, is updated at block 215 by averaging the last three frames with the current frame energy. Similarly, the short-term average of LSF vectors, $\overline{LSFs}$, is updated at block 220 by averaging the last three LSF frame vectors with the current LSF frame vector extracted by the parameter extractor at block 200. If the standard deviation σ is less than T_1 or the long-term mean of the pitch gain is greater than T_2, then a flag P_{flag} is set to one, otherwise P_{flag}

equals zero at block 225.

$$\text{If } \sigma < T_1 \text{ OR } P_{gain} > T_2, \text{ then } P_{flag} = 1, \text{ else } P_{flag} = 0.$$

In the preferred embodiment, $T_1 = 1.2$ and $T_2 = 0.7$. At block 230, a minimum energy buffer is updated with the minimum energy value over the last 128 frames. In other words, if the present energy level is less than the minimum energy level determined over the last 128 frames, then the value of the buffer is updated, otherwise the buffer value is unchanged.

If the frame count (i.e. current frame number) is less than a predetermined frame count N_I at block 235, where N_I is 32 in the preferred embodiment, an initialization routine is performed by blocks 240 - 255. At block 240 the average energy $\overline{E}$, and the long-term average noise spectrum $\overline{LSF_N}$ are calculated over the last N_I frames. The average energy $\overline{E}$ is the average of the energy of the last N_I frames. The initial value for $\overline{E}$, calculated at block 240, is:

$$\overline{E} = \frac{1}{N} \sum_{n=1}^{N} E$$

The long-term average noise spectrum $\overline{LSF_N}$ is the average of the LSF vectors of the last N_I frames. At block 245, if the instantaneous energy E extracted at block 200 is less than 15 dB, then the voicing decision is set to zero (block 255), otherwise the voicing decision is set one (block 250). The processing for the frame is then completed and the next frame is processed, beginning with block 200.

The initialization processing of blocks 240-255 initializes the processing over the last few frames. It is not critical to the operation of the present invention and may be skipped. The calculations of block 240 are required, however,

for the proper operation of the invention and should be performed, even if the voicing decisions of blocks 245-255 are skipped. Also, during initialization, the voicing decision could always be set to "1" without significantly impacting the performance of the present invention.

If the frame count is not less than N_I at block 235, then the first time through block 260 (Frame_Count = N_I), the long-term average noise energy $\overline{E_N}$ is initialized by subtracting 12 dB from the average energy $\overline{E}$:

$$\overline{E_N} = \overline{E} - 12dB$$

Next, at block 265, a spectral difference value SD_I is calculated using the normalized Itakura-Saito measure. The value SD_I is a measure of the difference between two spectra (the current frame spectra represented by R and E_π , and the background noise spectrum represented by $\vec{a}$. The Itakurass-Saito measure is a well-known algorithm in the speech processing art and is described in detail, for example, in *Discrete-Time Processing of Speech Signals*, Deller, John R., Proakis, John G. and Hansen, John H.L., 1987, pages 327-329, herein incorporated by reference. Specifically, SD_I is defined by the following equation:

$$SD_I = \frac{\vec{a}^T R \vec{a}}{E_\pi}$$

where E_π is the prediction error from linear prediction (LP) analysis of the current frame;

R is the auto-correlation matrix from the LP analysis of the current frame; and

$\vec{a}$ is a linear prediction filter describing the background noise

obtained from $\overline{LSF_N}$.

At block 270 the spectral differences SD_2 and SD_3 are calculated using a mean square error method according to the following equations:

$$SD_2 = \sum_{i=1}^{p} [\overline{LSF_S}(i) - \overline{LSF_N}(i)]^2$$

$$SD_3 = \sum_{i=1}^{p} [\overline{LSF_S}(i) - \overline{LSF}(i)]^2$$

Where $\overline{LSF_S}$ is the short-term average of LSF;

$\overline{LSF_N}$ is the long-term average noise spectrum; and

LSF is the current LSF extracted by the parameter extraction.

The long-term mean of SD_2 (sm_SD_2) in the preferred embodiment is updated at block 275 according to the following equation:

$$sm\_SD_2 = 0.4 * SD_2 + 0.6 * sm\_SD_2$$

Thus, the long term mean of SD_2 is a linear combination of the past long-term mean and the current SD_2 value.

The initial voicing decision, obtained in block 280, is denoted by I_{VD}. The value of I_{VD} is determined according to the following decision statements:

If $\overline{Es} \geq \overline{E}_N + X1$ dB

OR

$E > \overline{E}_N + X2$ dB

then IVD = 1;

If $\overline{E}s - \overline{E}_N < X3$ dB

AND

$sm\_SD_2 < T3$

AND

Frame_Count > 128

then IVD = 0 ; else IVD = 1;

If $E > 1/2\ (E^{-1} + E^{-2}) + X4$ dB

OR

SD1 > 1.5

then Ivd = 1 .

In the preferred embodiment, $X_1 = 1$, $X_2 = 3$, $X_3 = 2$, $X_4 = 7$, and $T_3 = 0.00012$.

The initial voicing decision is smoothed at block 285 to reflect the long term stationary nature of the speech signal. The smoothed voicing decision of the frame, the previous frame and the frame before the previous frame are denoted by S_{VD}^0, S_{VD}^{-1} and S_{VD}^{-2}, respectively. Both S_{VD}^{-1} and S_{VD}^{-2} are initialized to 1 and $S_{VD}^0 = I_{VD}$. A Boolean parameter F_{VD}^{-1} is initialized to 1 and a counter denoted by C_e is initialized to 0. The energy of the previous frame is denoted by E_{-1}. Thus, the smoothing stage is defined by:

$$
\begin{aligned}
&\textit{if } F_{VD}^{-1} = 1 \textit{ and } I_{VD} = 0 \textit{ and } S_{VD}^{-1} = 1 \textit{ and } S_{VD}^{-2} = 1 \\
&\quad S_{VD}^{0} = 1 \\
&\quad C_e = C_e + 1 \\
&\quad \textit{if } C_e \le T_4 \ \{ \\
&\qquad F_{VD}^{-1} = 1 \\
&\qquad \} \\
&\quad \textit{else } \{ \\
&\qquad F_{VD}^{-1} = 0 \\
&\qquad C_e = 0 \\
&\qquad \} \\
&\quad \} \\
&\textit{else} \\
&\quad F_{VD}^{-1} = 1
\end{aligned}
$$

Ce is reset to 0 if $S_{VD}^{-1} = 1$ and $S_{VD}^{-2} = 1$ and $I_{VD} = 1$.

If $P_{flag} = 1$, then $S^{0}_{VD} = 1$

If $E < 15$ dB, then $S^{0}_{VD} = 0$

In the preferred embodiment, $T_4 = 14$. The final value of S^{0}_{VD} represents the final voicing decision, with a value of "1" representing an active voice speech signal, and a value of "0" representing a non-active voice speech signal.

F_{SD} is a flag which indicates whether consecutive frames exhibit spectral stationarity (i.e., spectrum does not change dramatically from frame to frame). F_{SD} is set at block 290 according to the following where C_S is a counter initialized to 0.

If Frame_Count > 128 AND $SD_3 < T_5$
then
$C_S = C_S + 1$
else
$C_S = 0$;
If $C_S > N$
$F_{SD} = 1$
else
$F_{SD} = 0$.

In the preferred embodiment, T5 = 0.0005 and N = 20.

The running averages of the background noise characteristics are updated at the last stage of the VAD algorithm. At block 295 and 300, the following conditions are tested and the updating takes place only if these conditions are met:

If $\bar{E}_S < \bar{E}_N + 3$ AND $P_{flag} = 0$
then $E_N = \beta_{EN} * \bar{E}_N + (1 - \beta_{EN}) * [\max \text{ of } E \text{ AND } \bar{E}_S]$
AND $\overline{LSF}_N(i) = \beta_{LSF} * \overline{LSF}_N(i) + (1 - \beta_{LSF}) * LSF(i) \quad i = 1, \ldots p$
If Frame_Count > 128 AND
$\bar{E}_N <$ Min AND $F_{SD} = 1$ AND $P_{flag} = 0$
then
$\bar{E}_N =$ Min
else If Frame_Count > 128 AND $\bar{E}_N >$ Min + 10
then
$\bar{E}_N =$ Min.

Figure 3 illustrates a block diagram of one possible implementation of a VAD 400 according to the present invention. An extractor 402 extracts the required predetermined parameters, including a pitch lag and a pitch gain, from the incoming

speech signal 105. A calculator unit 404 performs the necessary calculations on the extracted parameters, as illustrated by the flowcharts in Figs. 2(A) and 2(B). A decision unit 406 then determines whether a current speech frame is an active voice or a non-active voice signal and outputs a voicing decision 140 (as shown in Fig. 1).

Those skilled in the art will appreciate that various adaptations and modifications of the just-described preferred embodiments can be configured without departing from the scope and spirit of the invention. Therefore, it is to be understood that within the scope of the appended claims, the invention may be practiced other than as specifically described herein.

CLAIMS

What Is Claimed Is:

1. In a speech communication system, a method for generating a frame voicing decision comprising the steps of:

(a) extracting a predetermined set of parameters, including a pitch gain and a pitch lag, from the incoming speech signal for each frame; and

(b) making a frame voicing decision according to the extracted predetermined set of parameters.

2. The method according to claim 1, wherein the predetermined set of parameters further comprises a full band energy and line spectral frequencies (LSF).

3. A method according to claim 2, wherein the step of making a frame voicing decision further comprises the steps of:

i. calculating a standard deviation σ of the pitch lag;
ii. calculating a long-term mean of pitch gain;
iii. calculating a short-term average of energy E, $\overline{E}_S$;
iv. calculating a short-term average of $\overline{LSF}_S$;
v. calculating an average energy $\overline{E}$; and
vi. calculating an average LSF value, $\overline{LSF}_N$.

4. A method according to claim 3, wherein the step of making a frame voicing decision further comprises the steps of:

i) calculating a spectral difference SD_1 using a normalized Itakura-Saito measure;

ii) calculating a spectral difference SD_2 using a mean square error method;

iii) calculating a spectral difference SD_3 using a mean square error method; and

iv) calculating a long-term mean of SD_2.

5. A method according to claim 4, wherein an initial frame voicing decision is made according to the calculated values.

6. A method according to claim 5, wherein the initial frame voicing decision is smoothed.

7. A method according to claim 6, wherein an initialization routine is performed for a predetermined number of initial frames, such that the voicing decision is set to active voice.

8. A voice activity detector (VAD) for making a voicing decision on an incoming speech signal frame, the VAD comprising:

an extractor for extracting a predetermined set of parameters, including a pitch gain and a pitch lag, from the incoming speech signal for each frame;

a calculator unit for calculating a set of predetermined values based on the extracted predetermined set of parameters: and

a decision unit for making a frame voicing decision according to the predetermined set of values.

9. The VAD according to claim 8, wherein the predetermined set of parameters further comprises a full band energy and line spectral frequencies (LSF).

10. The VAD according to claim 9, wherein the calculator unit calculates:

a standard deviation σ of the pitch lag;

a long-term mean of pitch gain;

a short-term average of energy E, $\bar{E}_S$;

a short-term average of LSF, $\overline{LSF_S}$;

an average energy $\bar{E}$; and

an average LSF value. $\overline{LSF_N}$.

11. The VAD according to claim 10, wherein the calculator unit further calculates:

a spectral difference SD, using a normalized-Itakura-Saito

measure;

a spectral difference SD_2 using a mean square error method;

a spectral difference SD_3 using a mean square error method; and

a long-term mean of SD_2.

12. The VAD according to claim 11, wherein the decision unit makes an initial frame voicing decision according to the values calculated by the calculation means

13. The VAD according to claim 12, wherein the initial frame voicing decision is smoothed.

14. A voice activity detection method for detecting voice activity in an incoming speech signal frame, the improvement comprising making a voicing decision based on a pitch lag and a pitch gain of the speech signal frame.

15. The voice activity detection method of claim 14, further comprising making the voicing decision based on a frame full band energy and a set of spectral parameters called Line Spectral Frequencies (LSF).

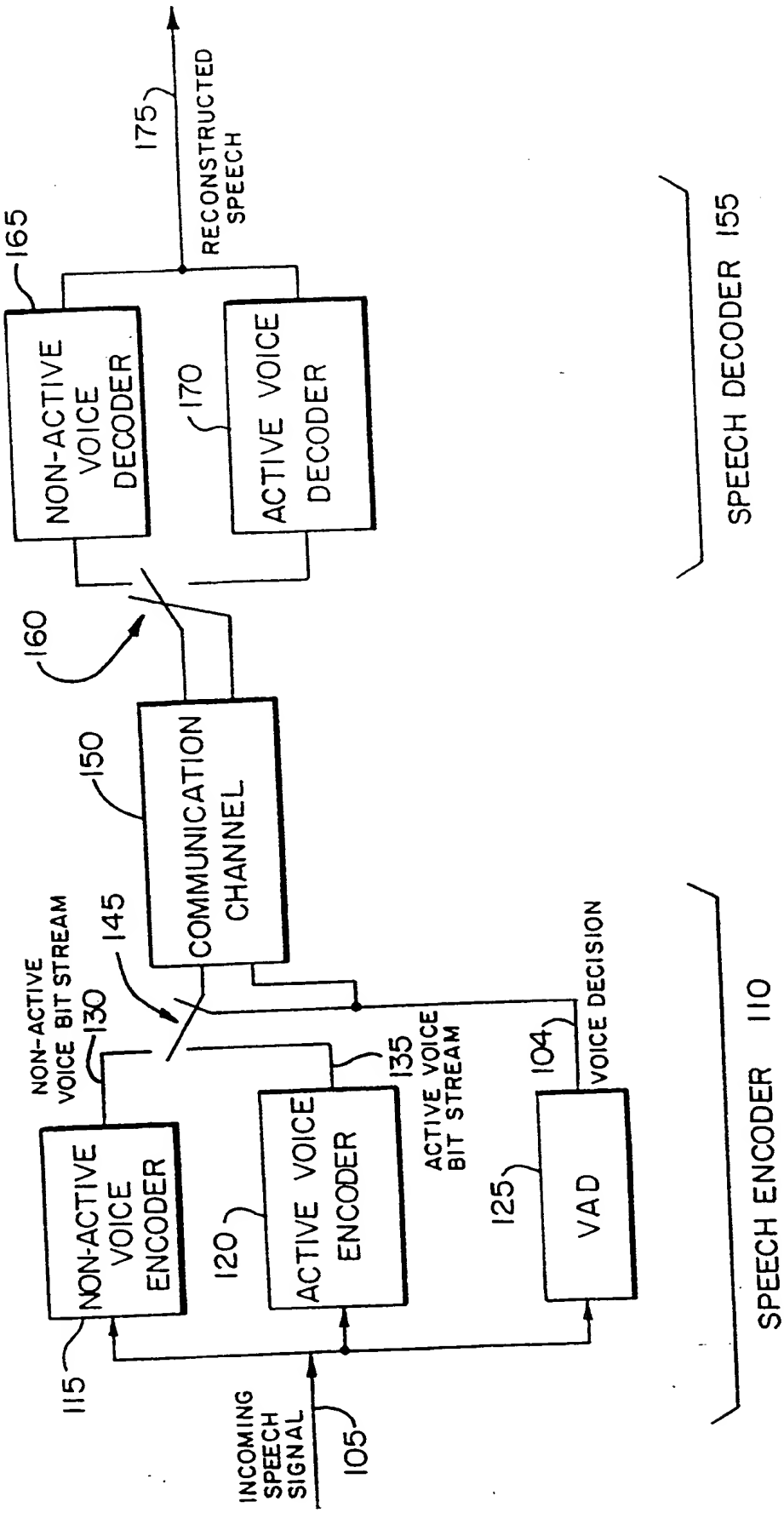


FIG. 1
PRIOR ART

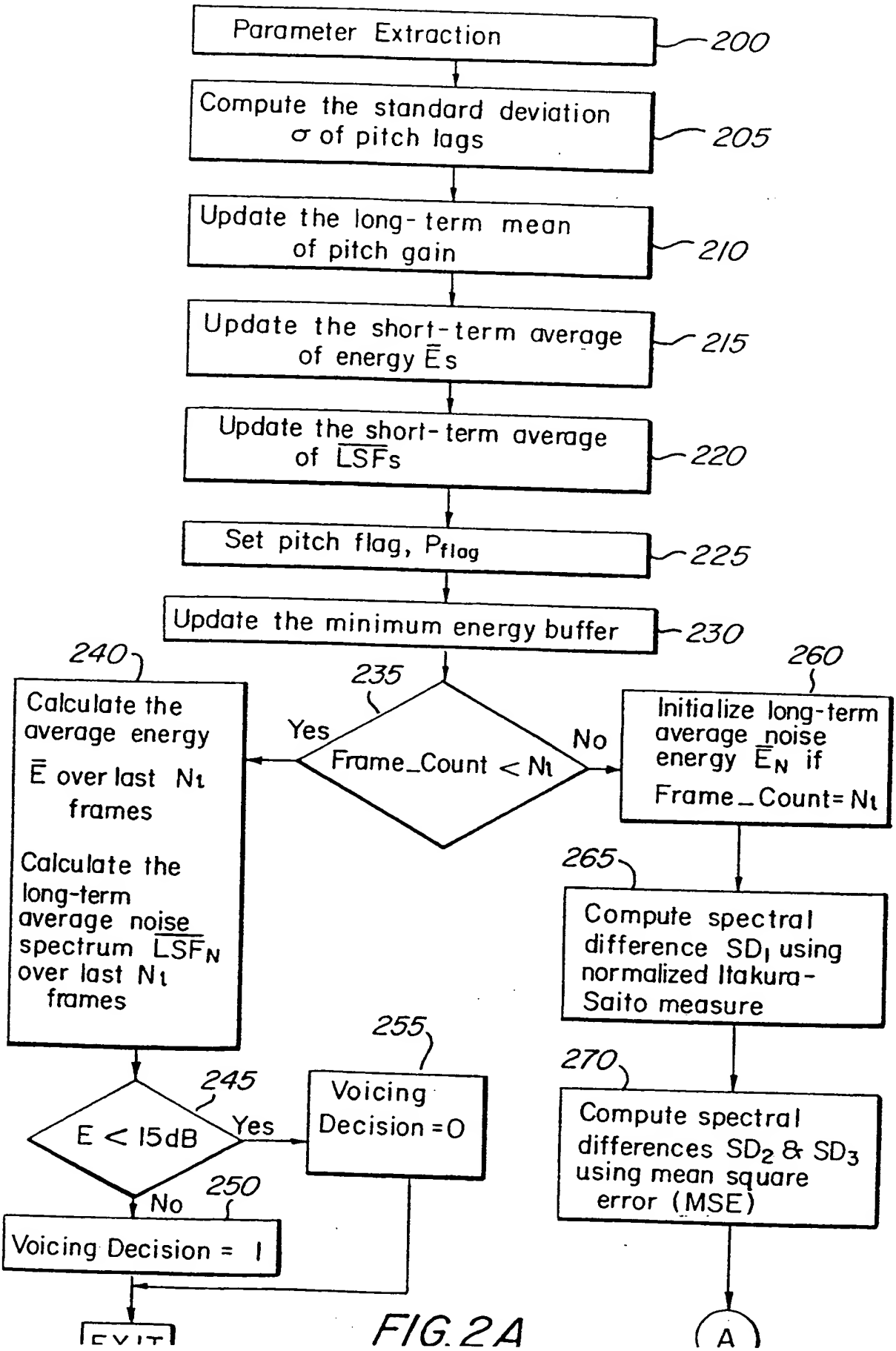

Parameter Extraction
200
Compute the standard deviation σ of pitch lags
205
Update the long-term mean of pitch gain
210
Update the short-term average of energy $\bar{E}_s$
215
Update the short-term average of $\overline{LSF}_s$
220
Set pitch flag, P_{flag}
225
Update the minimum energy buffer
230
235
Frame_Count < Nt
Yes
No
240
Calculate the average energy $\bar{E}$ over last Nt frames
Calculate the long-term average noise spectrum $\overline{LSF}_N$ over last Nt frames
260
Initialize long-term average noise energy $\bar{E}_N$ if Frame_Count = Nt
265
Compute spectral difference SD_1 using normalized Itakura-Saito measure
270
Compute spectral differences SD_2 & SD_3 using mean square error (MSE)
245
E < 15dB
Yes
No
255
Voicing Decision = 0
250
Voicing Decision = 1
EXIT
A

FIG. 2A

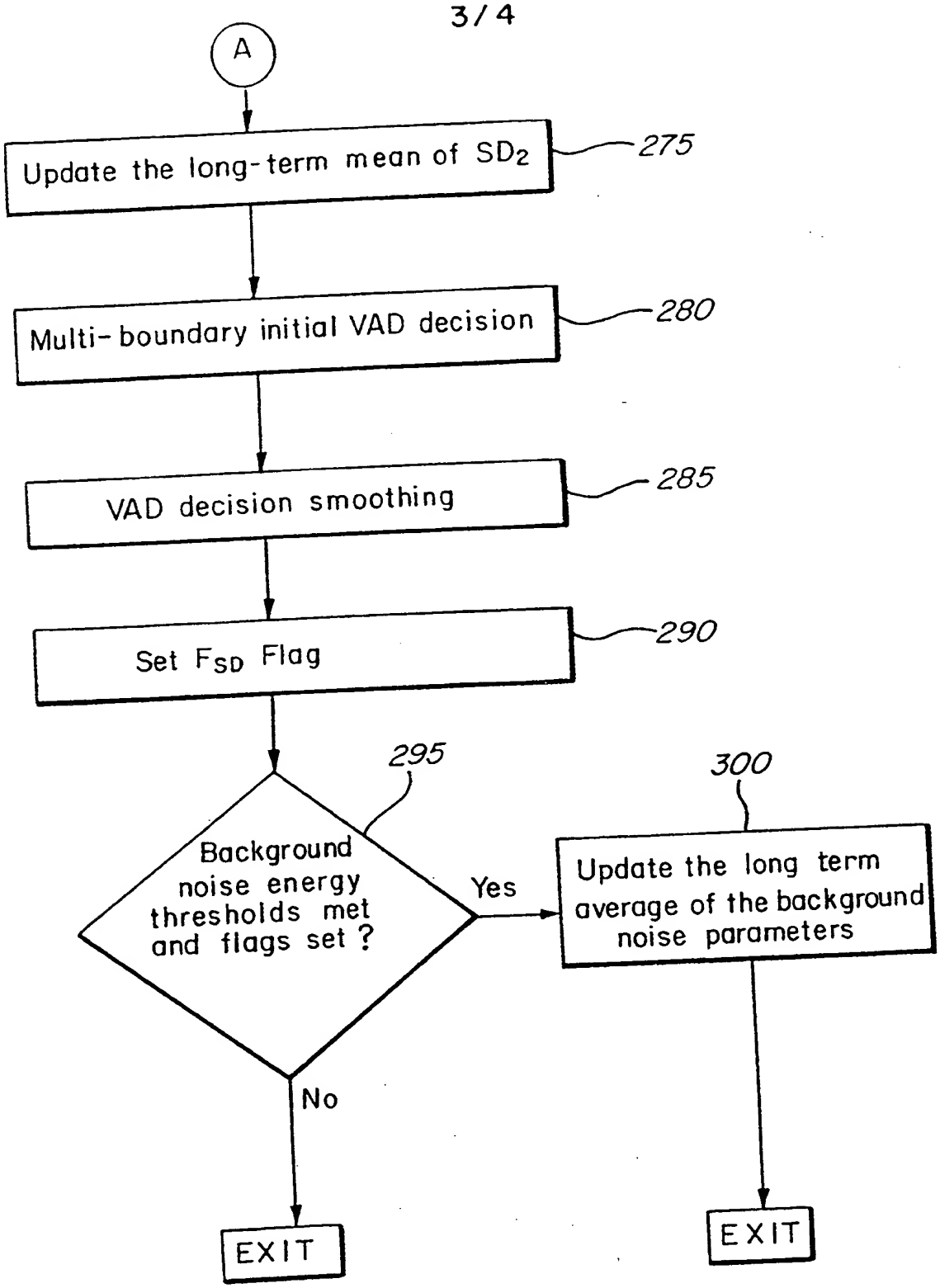
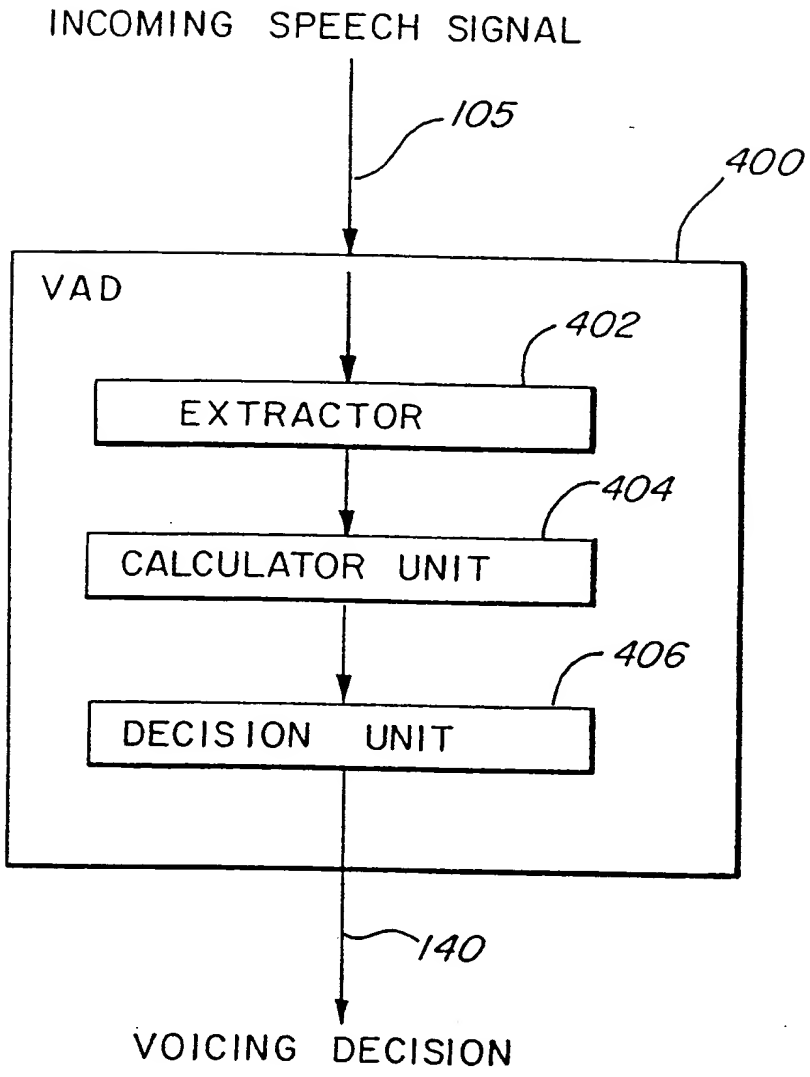

A
Update the long-term mean of SD2
275
Multi-boundary initial VAD decision
280
VAD decision smoothing
285
Set FSD Flag
290
295
Background noise energy thresholds met and flags set ?
Yes
No
300
Update the long term average of the background noise parameters
EXIT
EXIT

FIG. 2B

FIG. 3


INCOMING SPEECH SIGNAL
105
400
VAD
402
EXTRACTOR
404
CALCULATOR UNIT
406
DECISION UNIT
140
VOICING DECISION

INTERNATIONAL SEARCH REPORT

In ternational Application No
PCT/US 99/19806

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L11/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 784 311 A (NOKIA MOBILE PHONES LTD) 16 July 1997 (1997-07-16) abstract; figure 2 page 8, line 15 - line 20	1,8,14
Y		2-4, 9-11,15
Y	EP 0 785 419 A (ROCKWELL INTERNATIONAL CORP) 23 July 1997 (1997-07-23) cited in the application claims 1-14	2-4, 9-11,15

☐ Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 December 1999	11/01/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Van Doremalen, J

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/US 99/19806

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0784311 A	16-07-1997	FI 955947 A	13-06-1997
		AU 1067797 A	03-07-1997
		AU 1067897 A	03-07-1997
		EP 0790599 A	20-08-1997
		WO 9722116 A	19-06-1997
		WO 9722117 A	19-06-1997
		JP 9212195 A	15-08-1997
		JP 9204196 A	05-08-1997
		US 5839101 A	17-11-1998
		US 5963901 A	05-10-1999
EP 0785419 A	23-07-1997	US 5774849 A	30-06-1998
		JP 9198099 A	31-07-1997